图书情报工作
第62卷 第11期 2018年6月

专利文本主题建模中领域停用词自动选取研究*

俞琰[1,2] 赵乃瑄[1]

[1] 南京工业大学信息服务部 南京 210009 [2] 东南大学成贤学院电子与计算机学院 南京 211816

摘要：［目的/意义］针对专利文本主题建模中领域停用词自动选取尚未有充分研究的问题，提出一种新的领域停用词自动选取方法，用于专利文本主题模型分析，以提高专利主题模型的区分度与建模质量。［方法/过程］领域停用词本质上是信息比较少，在不同类别专利文本中区分度低的词。因此，引入辅助专利文本集，使用类别熵衡量词的分布情况，然后依据词的类别熵进行排序，选取类别熵最大的若干词作为领域停用词。［结果/结论］实验通过专利文本数据，验证了该方法的可行性与有效性，能够有效地提高专利主题模型的区分度。

关键词：专利文本 主题建模 领域停用词 自动选取

分类号：G202

DOI:10.13266/j.issn.0252-3116.2018.11.014

1 引言

有效的专利文本分析能够判断领域技术热点、识别领域核心技术，预测领域技术发展趋势，帮助研发人员从中获得启发与借鉴，从而缩短创新设计时间、节约创新设计经费。因此，专利文本分析具有重要的研究意义。传统的专利文本分析方法通常使用专利文本中的词语直接作为主题或概念，进而利用主题或概念建模，分析领域技术状况[1-5]。然而，专利作为一种被保护的文献，专利申请者为了扩大所申请专利的保护范围和提高专利授权的可能性，往往会使用一些模糊或者抽象的表达。因此，从专利文本所表达的潜在语义层面理解专利文本，才能得到更好的专利文本分析效果。不同于传统的文本分析方法，主题模型通过分析文本集合中词语共现的概率分布，挖掘文本隐含的语义信息，被广泛应用于文本分析之中，并取得较好的效果。随着主题模型的逐步完善，研究者开始尝试将主题模型应用于专利文本分析之中，以揭示专利文本深层次知识结构[6-14]。

虽然主题模型可有效地挖掘专利文本中隐含的语义信息，取得了较好的分析效果，然而，在主题模型学习过程中，学习得到的主题分布易向高频词倾斜。这些词通常是一些出现频率高但无实际意义的停用词，不能很好地刻画主题特征，如中文的“的”“是”，英文的“of”“the”等词。在生成主题模型的迭代过程中，这些词频繁出现在多个主题中，导致主题分布摇摆不停，两个主题分布相似性提高，不能明显区别各个主题，收敛速度变慢，对主题模型结果产生负面的影响[15]。为了避免这种情况，通常在构建专利文本主题模型之前借助停用词表，预先删除专利文本中的停用词。但是，这种方法并不能完全过滤掉表意性较差的词语。实际上，停用词不仅包括通用停用词，还包括领域停用词。前者是标准的共同领域停用词；后者为在特定领域中具有很少区分度的词。以专利文本分析为例，专利文本中常出现“方法”“包含”“发明”等词，这些词包含信息量少，区分度低，不能很好地表示专利文本的语义信息。一些专利文本主题模型研究采用手工方式或者词频和文本频次等相关方法选取领域停用词[9,13]。然而，基于词频或文本频次选取的领域停用词还可能包括一些有用的领域专利术语，如，在“3D 打印”专利文本主题分析中，词“打印机”大量出现在相关文本集中，具有较高的词频和文本频次，但其具有一定的研究价值，不能作为领域停用词简单删除。

中文专利文本中领域停用词表具有自身的特点。领域通用词通常具有主题无关性，在多个类别专利中均匀重复出现，如，“方法”“包含”“发明”等词在化学、

* 本文系教育部人文社科规划项目项目“大数据时代技能知识图谱构建研究”（项目编号：16YJAZH073）和国家社会科学基金一般规划项目“大数据时代支持创新设计的多维度多层次专利文本挖掘研究”（项目编号：17BTQ059）研究成果之一。

作者简介：俞琰（ORCID：0000-0002-9654-8614），副教授，博士，E-mail：yuyanyuyan2004@126.com；赵乃瑄，馆长，教授，博士。

收稿日期：2017-11-09 修回日期：2018-03-05 本文起止页码：120-126 本文责任编辑：王善军

机械、物理、电学等不同专利类别中均大量重复出现。相反,包含核心领域知识的专利术语则仅在某个类别中频繁出现,而在其他类别中出现频次较低,甚至不出现。因此,本文引入包含多类别专利文献的辅助专利文本集,提出类别熵的概念,衡量词的分布,以自动选取领域停用词,从而提高专利文本主题模型的区分度与建模质量。

本文结构如下:第 2 部分介绍相关研究工作;第 3 部分介绍用于专利文本分析的主题模型;第 4 部分为本文提出的领域停用词自动选取方法;第 5 部分为相关实验及实验结果分析;最后为本文结论。

2 相关研究

停用词被认为是无实际语义信息、无区分度的词。它们构成了文本数据的大部分,在文本分析过程中存在很大的干扰性,不仅携带较少的信息,还会对其他词语产生一定的抑制作用,很大程度上影响文本处理的效率和准确性。去除停用词被广泛使用于各种文本分析领域。如,W. Frakes 等[15]在信息检索的研究中认为自动索引阶段提早考虑消除出现频率过高的词语可以提高检索速度,减少检索存储空间并且不会降低检索结果的准确性。C. Silver[16]验证了基于支持向量机的文本分类器在去除停用词之后,准确率有所提高。官琴[17]选取百度停用词表、哈尔滨工业大学停用词表以及四川大学机器智能实验室停用词表,对不同聚类结果进行效果评估。研究结果表明停用词表对于文本聚类准确度有很大的影响,构建或选取适宜的停用词表极为重要。总的来说,停用词的选取对文本分析结果异常重要,去除停用词是文本预处理中十分重要的步骤。

停用词可分为通用停用词和领域停用词两大类。领域停用词因领域与数据集不同而不同。例如,词语“学习”在教育领域可能是一个领域停用词,但是在计算机科学领域可能不是一个领域停用词。

领域停用词已经被应用于人力资源管理[18]、生物信息、基因本体[19]、信息检索[20]和电子商务[21]等领域之中。通常,选取领域停用词通过手工完成,而自动选取领域停用词可根据实际处理文本集的不同而自动构造合适的领域停用词,灵活性强,更具潜力。但是如何设计高效准确的自动选取领域停用词算法也是具有挑战性的任务。通常采用词频或文本频次进行领域停用词的自动选取。基于词频的领域停用词自动选取理论依据是若一个词在文本集中大量出现,则认为该词是停用词。文本频次则计算文本集中出现某个词的文本数来表示。其理论假设是当一个词在大量文本中出现时,该词不具有较强的文本区分能力,可被认为是领域停用词。

此外,一些研究者尝试采取其他一些方法自动选取领域停用词。例如,T. W. Lo 等[22]针对信息检索,提出一种基于词语的随机抽样抽取方法,并提出最有效的停用词表是经典的停用词表和新方法自动抽取的停用词表的融合。L. Hao 等[23]提出 x2-统计方法,产生家具种类查询的领域停用词,以加速电子商务网站信息检索过程。M. P. Sinka 和 D. W. Corne[24]提出单词熵,使用聚类和随机检索算法优化,自动选取领域停用词。M. Jungiewicz 和 M. Lopuszynski[25]基于观察:每个文本的停用词的出现次数的分布通常遵循一个典型的随机变量模型(如,泊松分布),开发了一个非监督方法自动产生领域停用词。M. Makrehchi 和 M. S. Kamel[26]假设停用词具有最小信息和预测能力,提出后向过滤级别性能和数据稀疏索引的概念,从一个标记集合中自动产生领域停用词,用于文本分类。顾益军等[27]分别计算词条在语料库中各个句子内发生的概率和包含该词的句子在语料库中的概率,在词基础上计算联合熵,依据联合熵选取领域停用词。巩政和关高娃[28]采用联合熵算法初步确定蒙古文停用词,接着从初步确定的蒙古文停用词中去掉蒙古文实体名词及同形异义词,再通过对英文停用词和蒙古文停用词的词性比较,确定蒙古文停用词表。珠杰和李天瑞[29]结合现有停用词的处理技术,研究基于统计的藏文停用词选取方法,通过实验分析了词项频率、文档频率、熵等方法的藏文停用词选取情况,提出了藏文叙词、特殊动词和自动处理方法相结合的藏文停用词选取方法。专利中领域停用词有其自身特点,这些方法并不适用于专利文本处理。

3 LDA 主题模型

LDA(Latent Dirichlet Allocation)模型[30]是一种常用的主题模型,由于其参数简单,不产生过拟合现象,逐渐成为主题模型的研究热点。本文使用 LDA 模型对专利文本进行建模。LDA 是一个三层贝叶斯概率模型,由词、主题和文本三层构成。该模型假设每个文本包含若干隐含主题,每个主题包含特定的词。文本和词间的关系通过隐含主题体现。隐含主题之间是相互独立的,这些主题被文本集中所有文本所共享,而每个文本有一个特定的主题分布。模型通常采用 Gibbs 采

样推理方法估计主题的后验分布,计算如公式(1)[24]所示:

$$p(z_{ij}=k \mid z^{-ij},w,\alpha,\beta) \propto \frac{n_{i(\cdot)k}^{-ij}+\beta}{n_{(\cdot)(\cdot)k}^{-ij}+V\beta} \times \frac{n_{(\cdot)jk}^{-ij}+\alpha}{n_{(\cdot)j(\cdot)}^{-ij}+K\alpha} \quad (1)$$

其中,z_{ij}表示文本 d_j 中词 w_i 的主题变量;$-ij$ 表示排除文本 d_j 中的词 w_i;n_{ijk}表示文档 d_j 中的词 w_i 分配给主题 k 的次数;$(\cdot)$表示对应维度(词语、文本、主题)所有次数之和,β 表示词的 Dirichlet 先验分布,α 表示主题的 Dirichlet 先验分布,K 表示主题数,V 表示集合中总的词语数。一旦获得每个文本中每个词的主题,就可以得到 LDA 模型中 θ 和 φ 的后验估计值,计算如公式(2)[24]和(3)[24]所示:

$$\theta_{jk}=\frac{n_{(\cdot)jk}+\alpha}{n_{(\cdot)j(\cdot)}+K\alpha} \quad (2)$$

$$\varphi_{ki}=\frac{n_{i(\cdot)k}+\beta}{n_{(\cdot)(\cdot)k}+V\beta} \quad (3)$$

其中,θ_{jk}表示文本 d_j 包含主题 k 的概率;φ_{ki}表示主题 k 中词语 w_i 的概率。

4 领域停用词自动选取

相较于专利术语,专利中的领域停用词通常具有类别无关性,在各种类别中反复均匀出现。相反,包含核心领域知识的专利术语则仅在某个类别中频繁出现,而在其他类别中出现频次较低,甚至不出现。因此,本文引入包含多个类别的辅助专利文本集,以识别专利文本中的领域停用词。专利中的领域停用词在各类别间以及某一类别内通常均匀出现。而词在类别间和类别内分布情况可以使用信息熵来衡量。信息熵是信息论中重要的概念,用来度量信息的不确定程度。词在文本中出现具有一定的不确定性,当词在文本间分布不均匀时,词提供给文本集的信息量越大,说明它区分文本的能力越强,这种不均匀性可以用词的信息熵来度量,衡量词在文本集中分布情况。

具体地,本文引入包含 $c_1, c_2, \cdots, c_m$ 类别的辅助专利文本集,每个类别包含若干个相关专利文本,将词 w_k 在不同专利类别间的分布称为类别间信息熵(Entropy between Categories, EBC),计算公式如下:

$$EBC(w_i) = -\sum_{t=1}^{m}\frac{df(w_i,c_t)}{df(w_i)} \times lb\frac{df(w_i,c_t)}{df(w_i)} \quad (4)$$

其中,$EBC(w_i)$表示词 w_i 的类别间信息熵;$df(w_i,c_t)$表示词 w_i 在类别 c_t 中的文档频次;$df(w_i)=\sum_{t=1}^{m} df(w_i,c_t)$,表示词 w_i 在辅助专利文本集中的文档频次。由定义可知,当词只出现在单个类别的文本中时,类别间信息熵最小;当词在所有类别中均匀分布时,类别间信息熵达到最大值。

类似地,本文将词 w_i 在类别 c_t 内的分布使用类别内信息熵(Entropy in Category, EIC)衡量,其计算公式如下:

$$EIC(w_i,c_t) = -\sum_{j=1}^{|c_t|}\frac{tf(w_i,d_j)}{tf(w_i,c_t)} \times lb\frac{tf(w_i,d_j)}{tf(w_i,c_t)} \quad (5)$$

其中,$EIC(w_i, c_t)$表示词 w_i 在类别 c_t 内的类别内信息熵;$tf(w_i,d_j)$表示词 w_i 在领域 c_t 的文档 d_j 中的词频;$tf(w_i,c_t)=\sum_{d_j \in c_t} tf(w_i,d_j)$,为 w_i 在类别 c_t 中的总词频,$|c_t|$表示类别 c_t 中包含的文档数。由定义可知,词在类别内分布越均匀,类别内信息熵值越大;反之,词在类别内分布越不均匀,类别内信息熵值越小。

结合类别间信息熵 EBC 和类别内信息熵 EIC,形成词的类别熵 EC,用于衡量词在各个类别的分布情况,计算公式如下:

$$E(w_i)=EBC(w_i) \times \sum_{t=1}^{m} EIC(w_i,c_t) \quad (6)$$

由定义可知,类别熵 E 越大,表明词在各领域内、领域间分布越均匀,越可能是专利文本中的领域停用词。

5 实验

5.1 数据集与实验设置

为了验证提出模型的有效性,本部分分别选取 3D 打印与智能语音相关专利文本进行实验。3D 打印是一项新兴制造技术,因其在某种程度上颠覆了传统制造业的生产方式,带来制造业数字化和智能化的革命,受到各国学术界和产业界的广泛关注,近年来取得快速发展。智能语音是人机交互模式的新选择。借助于移动互联网、机器学习领域中深度学习技术以及大数据语料库的积累,智能语音技术的实用化发展突飞猛进,在电信、金融、汽车电子、家电、教育、玩具、智能手机、移动互联网等领域已得到广泛应用。实验基于中国国家知识产权局专利数据库,分别以“3D 打印 or 快速成型 or 增材制造 or 三维打印 or 增量制造 or 添加制造 or 智能制造 or 数字化制造”和“智能语音 or 语音识别 or 语音合成 or 自然语言理解 or 语音交互 or 语音技术 or 语音控制”作为检索式,检索 2013 年至 2017 年相关专利文献(检索日期为 2017 年 8 月 1 日)。通过数据抓取、清洗、去重后,最终分别将 7 790 条 3D

打印、5 272 条智能语音中国发明公开专利标题和摘要作为待分析的目标专利文本集。

此外，实验根据专利 IPC 分类号，分别从 A－H 分类号中随机抽取 2 000 条中国发明公开专利文献标题和摘要作为辅助专利文本集，数据集统计信息如表 1 所示：

表 1 数据集基本信息

数据集类型	领域	#去重后文本
目标专利文本集	3D 打印	7 790
	智能语音	5 272
辅助专利文本集	A 人类生活必需(农、轻、医)	2 000
	B 作业;运输	2 000
	C 化学;冶金	2 000
	D 纺织;造纸	2 000
	E 固定建筑物(建筑、采矿)	2 000
	F 机械工程;照明;加热;武器;爆破	2 000
	G 物理	2 000
	H 电学	2 000

实验首先对目标专利文本集合采用中国科学院计算研究所的 ICTCLAS 分词系统进行分词，采用哈尔滨工业大学停用词列表对专利文本移除通用停用词。在 LDA 建模过程中，参数估计采用 Gibbs 采样算法。主题模型设置 $a=50/K$、$\beta=0.01$，Gibbs 采样迭代次数参数为 2 000，保存迭代参数为 1 000。主题数 K 的选取通过计算基本专利主题模型的困惑度选取最优值，采用五折交叉验证。根据计算，实验设定 3D 打印数据集的主题数 $K=15$、智能打印数据集的主题数 K＝10。

5.2 评估标准

实验借助平均 KL 距离指标定量描述主题的区分度。平均 KL 距离常用来衡量两个概率分布的距离。平均 KL 距离 avg_KL 的定义如下：

$$avg_KL = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K} KL(\varphi_{i} | \varphi_{j})}{K^2} \tag{7}$$

其中 $KL(\varphi_i \parallel \varphi_j) = \sum_{v=1}^{V}\varphi_{iv}\log\frac{\varphi_{iv}}{\varphi_{jv}}$。由于 KL 距离是不对称的，但是 φ_i 和 φ_j 相似性度量是对称的，故将公式进行调整，采用对称的 Jensen-Shannon 距离度量 2 个主题距离，具体计算公式如：

$$JS(\varphi_i,\varphi_j) = \frac{KL(\varphi_i,\varphi_j)+KL(\varphi_j,\varphi_i)}{2} \tag{8}$$

此时，平均 KL 距离衡量的是包含所有词的主题之间的距离，由于移除不同的停用词之后形成的专利文本集中包含的单词数不同，为了有效比较，形成新平均 KL 距离指标 avg_KL′衡量主题间单词的平均距离，计算公式如下：

$$avg_KL' = lb(avg_KL/V) \tag{9}$$

其中 V 为专利文本集中包含的单词数。此时，avg_KL' 值越大，表明主题与主题之间的距离越远，主题的可区分性越高。

5.3 实验结果

5.3.1 领域停用词选取阈值确定 实验使用 ICTCLAS 分词系统对辅助专利文本集进行分词，采用哈尔滨工业大学停用词列表移除通用停用词，使用第 4 部分提出的类别熵，计算词的类别熵。表 2 为类别熵分值最大的前 20 个词。由表 2 可见，类别熵值最高的这些词通常在各专利文献类别中均会出现，与具体专利主题分析中的专利术语无关，包含语义信息较少，可以作为领域停用词。

表 2 类别熵值最大的前 20 个词

序号	词	类别熵	序号	词	类别熵
1	种	257.12	11	后	234.97
2	发明	256.83	12	提供	231.79
3	包括	256.83	13	内	229.87
4	中	254.55	14	提高	228.42
5	公开	243.32	15	应	226.90
6	上	240.73	16	技术	225.42
7	下	239.96	17	简单	224.81
8	述	239.91	18	采用	224.60
9	方法	235.17	19	用于	223.30
10	涉及	234.98	20	之间	223.06

将领域停用词选取的类别熵阈值分别设为 120、130、140、150、160 和 170。实验分别将大于类别熵阈值的词作为领域停用词，在目标专利文本集中去除领域停用词，建立专利文本主题模型，实验结果如图 1 所示。由图 1 可见，当阈值为 150 时，3D 打印数据集和智能打印数据集中 avg_KL′值最大，专利文本主题模型的区分度最大。因此，本文在后续实验中选取类别熵阈值为 150。

5.3.2 不同领域停用词方法比较 为比较不同领域停用词选取方法专利文本主题建模效果，实验分别依据以下方法自动选取领域停用词：①TF：依据词在目标专利文本集中出现的词频，选取词频最高的若干词作为领域停用词；②DF：依据词在目标专利文本集中的文本频次，选取文本频次最高的若干词作为领域停用词；③EC：依据本文第 4 部分提出的类别熵选取领域停用词。

依据以上的三种方法选取领域停用词，生成不同

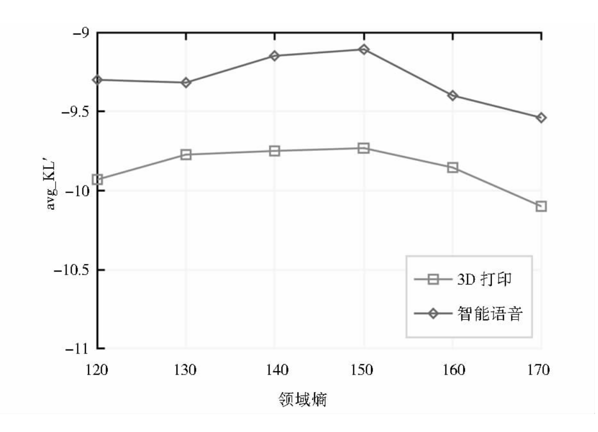


图 1　领域停用词阈值对区分度的影响

的专利文本集,分别学习专利主题模型,对应的专利主题模型分别称为 TF-LDA、DF-LDA 和 EC-LDA。为了便于比较,实验将没有移除任何停用词(包括通用停用词)和仅仅移除通用停用词的主题模型作为基本模型进行比较,分别称为 LDA 和 Gen-LDA。

由表 3 可见,首先,在 3D 打印数据集和智能打印数据集中,Gen_LDA 模型的平均 KL 距离均大于 LDA,这表明在进行专利文本主题模型建模时,利用通用停用词表移除通用停用词,能够改善主题模型性能,增加主题模型的区分度。其次,在两个专利文本集中,TF-LDA 模型的平均 KL 距离大于标准方法 LDA。这表明当使用词频作为标准,移除适当数量的高词频词,比起不移除任何停用词,能够改善主题模型的性能,增加主题模型的区分度。再次,DF-LDA 的平均 KL 距离大于 TF-LDA。DF-LDA 模型依据词的文本频次选取领域停用词,这表明考虑文档频次方法在专利文本主题建模时优于词频方法,通过词语所在文本的数目反映词语的可区分度与重要性,从而产生更好的主题建模效果。总体而言,DF-LDA 的平均距离大于 LDA 模型,略小于 Gen-LDA 模型。可能的解释是一些出现在若干专利文本中的、能够反映领域内容的词语被作为领域停用词

表 3　主题区分度比较

数据集	方法	Avg_K'
3D 打印	LDA	-11.15
	Gen_LDA	-11.01
	TF-LDA	-11.07
	DF-LDA	-11.04
	EC-LDA	-9.73
智能语音	LDA	-10.66
	Gen_LDA	-10.52
	TF-LDA	-10.59
	DF-LDA	-10.53
	EC-LDA	-9.11

被移除,影响了专利文本主题模型的区分度。最后,本文提出的 EC-LDA 模型的平均 KL 距离在两个数据集中均最大,这表明引入辅助专利文本集,利用类别熵度量词的分布程度,能够更加准确地度量词语的分布,表明其包含的信息量,从而取得最佳性能。

表 4 列出 3D 打印数据集和智能语音数据集中前 20 个最高 TF 和 DF 的词,其中领域停用词使用粗体表示。TF 选取目标专利集合中的高频词作为领域停用词,DF 依据词在目标专利集合中出现的文本频次选取领域停用词。由表 4 可见,在两个数据集中,除了一些信息量很少的词,如"发明""包括"等之外,也包括一些"打印""3D""语音""模块"等领域术语,简单删除这些词,影响了最终的专利主题模型的建模效果。

表 4　前 20 最高 TF 和 DF 词

序号	3D 打印		智能语音	
	TF	DF	TF	DF
1	打印	发明	语音	语音
2	3D	打印	模块	发明
3	发明	3D	识别	包括
4	述	包括	述	识别
5	方法	方法	控制	方法
6	包括	公开	方法	述
7	材料	述	发明	控制
8	装置	中	信息	系统
9	打印机	上	系统	公开
10	上	材料	装置	装置
11	制备	技术	包括	中
12	三维	打印机	用户	用户
13	中	制备	信号	信息
14	结构	装置	中	模块
15	模型	结构	智能	提供
16	连接	三维	输入	接收
17	用于	成型	连接	输入
18	技术	固定	指令	技术
19	成型	制造	用于	连接
20	喷头	模型	单元	信号

5.3.3　专利主题词比较　最后,为了得到直观效果,实验分别给出使用 Gen-LDA 和 EC-LDA 模型在 3D 打印专利文本和在智能语音专利文本模型每个主题中的前 5 个词。Gen-LDA 模型为通常采用的方法,仅仅使用停用词表去除通用停用词,而 EC-LDA 则在去除通用停用词基础上,使用本文第 4 部分提出的自动选取领域停用词的方法,移除领域停用词。结果如表 5、6 所示,领域停用词使用粗体表示。由表 5、表 6 可见,在 Gen-LDA 主题模型中,常出现一些表意性较差的词,如

俞琰，赵乃瑄. 专利文本主题建模中领域停用词自动选取研究[J]. 图书情报工作,2018,62(11):120－126.

"发明""方法""技术""包括"领域停用词等。相比于传统 Gen-LDA 方法,EC-LDA 模型根据类别熵,自动选取信息量小和区分度低的词作为领域停用词,使得主题一致性较强,更易于理解。

表 5 3D 打印专利文本集中主题词比较

主题	Gen-LDA	EC-LDA
0	制备 方法 粉末 中 得到	患者 制作方法 牙齿 手术 个性化
1	系统 控制 装置 模块 包括	混凝土 墙体 建筑 框架 墙体
2	发明 技术 3D 涉及 领域	凹槽 螺纹 结构 支撑 滑轨
3	表面 结构 上 形成 方法	粉末 陶瓷 金属 合金 制备
4	机构 上 装置 打印机 平台	移动 驱动 组件 电极 导轨
5	用于 包括 发明 多个 部分	模具 工艺 零件 一体化 蜡
6	方法 支架 固化 制备 发明	图像 参数 计算机 软件 扫描
7	成型 材料 过程 提高 发明	复合材料 原料 强度 改性 纳米
8	打印 3D 发明 方法 提高	膜 电极 基板 导电 芯片
9	装置 喷头 打印机 加热 挤出	传感器 温度 控制器 信号 电路
10	三维 模型 方法 进行 数据	腔 孔 通道 壳体 密封
11	连接 固定 设置 结构 安装	巧克力 食品 蛋糕 原材料 色彩
12	制造 方法 加工 激光 金属	树脂 光源 液态 胶 快速成型
13	定位 设计 患者 制作方法 发明	进料 喷嘴 耗材 螺杆 供料
14	材料 制备 重量 复合材料 具有	支架 生物 修复 纤维 细胞

表 6 智能语音专利文本中主题词比较

主题	Gen-LDA	EC-LDA
0	语音 识别 输入 用于 发明	数据 音频 识别 语音 生成
1	系统 交互 智能 机器人 基于 发明	特征 模型 合成 训练 解码
2	模块 系统 语音 技术 计算机	模块 电路 无线 传感器 通信
3	信号 连接 电路 述 发明	计算机 汉语 方案 输入 程序
4	控制 语音 指令 发明 用于	语音 装置 检测 判断 车载
5	数据 中 方法 文本 音频 包括	信息 移动 服务器 发送 播放
6	语音 方法 特征 模型 进行	语音 信号 输入 输出 声音
7	发明 进行 检测 时 识别 方法	系统 交互 机器人 智能 平台
8	信息 用户 语音 方法 述	装置 安装 电子 开关 显示屏
9	述 装置 上 智能 包括	语音 指令 命令 智能家居 遥控器

6 总结

针对专利文本主题建模中领域停用词自动选取尚未有相关研究,以及目前其他文本分析中常见领域停用词自动选取可能存在的问题,本文根据专利文献的特点,引入辅助专利文本集,提出类别熵,衡量词的分布情况,以自动选取领域停用词,用于专利文本主题模型分析,以提高专利主题模型的区分度与建模质量。通过实验,表明相比于传统的基于词频和文本频次的方法,使用本文提出的类别熵方法能够更好地衡量词的分布特征,更好地构建专利主题模型,增加专利主题之间的距离,增加可区分度。

参考文献:

[1] YOON B, PARK Y. A text-mining-based patent network: analytical tool for high-technology trend[J]. Journal of high technology management research, 2004, 15(1):37－50.

[2] 郭炜强, 戴天, 文贵华. 基于领域知识的专利自动分类[J]. 计算机工程, 2005, 31(23):52－54.

[3] KIM M, PARK Y, YOON J. Generating patent development maps for technology monitoring using semantic patent-topic analysis[J]. Computers & industrial engineering, 2016, 98(1):289－299.

[4] 高利丹, 肖国华, 张娴,等. 共现分析在专利地图中的应用研究[J]. 现代情报, 2009, 29(7):36－39.

[5] 张杰, 刘美佳, 翟东升. 基于专利共词分析的 RFID 领域技术主题研究[J]. 科技管理研究, 2013, 33(10):129－132.

[6] TANG J, WANG B, YANG Y, et al. PatentMiner: topic-driven patent analysis and mining[C]// ACM SIGKDD international conference on knowledge discovery and data mining. New York:ACM, 2012:1366－1374.

[7] WANG B, LIU S, DING K, et al. Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: a case study in LTE technology[J]. Scientometrics, 2014, 101(1):685－704.

[8] CHEN H, ZHANG G, LU J, et al. A fuzzy approach for measuring development of topics in patents using Latent Dirichlet Allocation[C]// IEEE international conference on fuzzy systems. Piscataway:IEEE, 2015:1116－1116.

[9] SUOMINEN A, TOIVANEN H, SEPPÄNEN M. Firms' knowledge profiles: Mapping patent data with unsupervised learning[J]. Technological forecasting & social change, 2017, 115(1):131－142.

[10] 范宇, 符红光, 文奕. 基于 LDA 模型的专利信息聚类技术[J]. 计算机应用, 2013, 33(S1):87－89.

[11] 王博, 刘盛博, 丁堃,等. 基于 LDA 主题模型的专利内容分析方法[J]. 科研管理, 2015, 36(3):111－117.

[12] 吴菲菲, 张亚茹, 黄鲁成,等. 基于 AToT 模型的技术主题多维动态演化分析——以石墨烯技术为例[J]. 图书情报工作, 2017,61(5):95－102.

[13] 廖列法, 勒孚刚. 基于 LDA 模型和分类号的专利技术演化研究[J]. 现代情报, 2017, 37(5):13－18.

[14] 陈亮, 张静, 张海超,等. 层次主题模型在技术演化分析上的应用研究[J]. 图书情报工作, 2017,61(5):103－108.

[15] FRAKES W B, BAEZA-YATES R. Information retrieval: data structures and algorithms[M]. 出版地: Prentice-Hall, Inc., 1992.

[16] SILVA C, RIBEIRO B. The importance of stop word removal on recall values in text categorization[C]// International joint conference on neural networks. Piscataway:IEEE, 2003:1661－1666.

[17] 官琴, 邓三鸿, 王昊. 中文文本聚类常用停用词表对比研究[J]. 现代图书情报技术, 2017 (3):72－80.

[18] CROW D, DESANTO J. A hybrid approach to concept extraction

and recognition-based matching in the domain of human resources [C]// IEEE international conference on TOOLS with Artificial Intelligence. Piscataway: IEEE, 2004:535-541.

[19] SEKI K, MOSTAFA J. An application of text categorization methods to gene ontology annotation[C]// International Conference on Research and Development in Information Retrieval. New York: ACM, 2005:138-145.

[20] TONG S, LERNER U, SINGHAL A, et al. Locating meaningful stopwords or stop-phrases in keyword-based retrieval systems [EB/OL]. [2018-04-06]. http://www.google.com/patents/US8626787.

[21] WHITE B J. Impact of domain-specific stop-sord lists on ECommerce website search performance[J]. Journal of strategic e-commerce, 2007, 5(2):83-101.

[22] LO T W, HE B, OUNIS I. Automatically building a stopword list for an information retrieval system[J]. Journal of digital information management, 2005, 3(1):3-8.

[23] HAO L, HAO L. Automatic identification of stop words in Chinese text classification[C]// International conference on computer science and software engineering. Piscataway: IEEE Computer Society, 2008:718-722.

[24] SINKA M P, CORNE D W. Evolving better stoplists for document clustering and Web intelligence[C]//T Design and application of hybrid intelligent systems, His03, thhe third international conference on hybrid intelligent system. New York: ACM, 2008:1015-1023.

[25] JUNGIEWICZ M, ŁOPUSZYŃSKI M. Unsupervised keyword extraction from polish legal texts[C]// International conference on natural language processing. Berlin: Springer International Publishing, 2014:65-70.

[26] MAKREHCHI M, KAMEL M S. Extracting domain-specific stop words for text classifiers[J]. 期刊名, 2017, 21(1):39-62.

[27] 顾益军, 樊孝忠, 王建华, 等. 中文停用词表的自动选取[J]. 北京理工大学学报, 2005, 25(4):337-340.

[28] 巩政, 关高娃. 蒙古文停用词和英文停用词比较研究[J]. 中文信息学报, 2011, 25(4):35-38.

[29] 珠杰, 李天瑞. 藏文停用词选取与自动处理方法研究[J]. 中文信息学报, 2015, 29(2):125-132.

[30] BLEI D M, NG A Y, Jordan M I. Latent Dirichlet allocation[J]. Journal of machine learning research, 2003, 3(1):993-1022.

作者贡献说明:

俞琰:提出研究思路,设计研究方案,进行实验,撰写论文;

赵乃瑄:采集、清洗和分析数据,修改论文。

Automatic Selection of Domain-Specific Stopwords in Topic Model of Patent Text

Yu Yan[1,2] Zhao Nianxuan[1]

[1] Information Service Department, Nanjing Tech University, Nanjing 210009

[2] Computer Science department, Southeast University Chengxian College, Nanjing 211816

Abstract: [**Purpose/significance**] Because the research that automatic selection of domain-specific stopwords in topic model of patent text is insufficient, this paper proposes a new method of automatic selection of domain-specific stopwords, for patent text topic model analysis, in order to improve the differentiation and modeling quality of the patent topic model. [**Method/process**] In essence, domain-specific stopwords are less important words which contain relatively less information, such words are poorly differentiated in different kinds of patent. Therefore, this paper introduced the auxiliary multi-category patent text dataset and measured the distributions of words through the category entropy. Then, according to the category entropy of words. It chose some words that have the maximum category entropy as the domain-specific stopwords. [**Result/conclusion**] Experimental results show the feasibility and validity of the method proposed in this paper, which can improve the differentiation and quality of topic model for patent text analysis.

Keywords: patent text topic model domain-specific stopword automatic selection